NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT 1103-40-11625

1 July 1959

NBS REPORT 6513

ON THE "SYNTHETIC RECORD" PROBLEM
(Estimation of the variance)

by

Joan Raup Rosenblatt

Statistical Engineering Laboratory

Technical Report No. 1
to
Water Resources Division
U. S. Geological Survey
Department of Interior

NBS

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

~~Project 5150 (USGS)~~
Technical Note No. 1
1 July 1959

On the "Synthetic Record" Problem

(Estimation of the Variance)

Joan Raup Rosenblatt
National Bureau of Standards

1. Introduction

The "synthetic record" problem (my terminology) arises in the following way. Discharge records are obtained for two streams in the same geographical area, one record being longer than the other. It is desired to find conditions under which the data in the longer record can be used to improve the estimates of mean and variance of the discharge in the stream for which there is a short record. Knowledge of these conditions would contribute to (i) evaluation of the "quality" of estimated parameters of discharge distributions and (ii) determination of criteria for the establishment and continuation of stream-gaging stations.

The term "synthetic record" is used because it literally describes one feature of standard hydrologic practice. It is customary to use the data from a long record of discharges to obtain estimates of the discharges in another stream for the corresponding dates, and to publish the resulting record.

2. Statistical Model and Assumptions

It is assumed that the simultaneous discharges (X, Y) from two streams have a joint normal distribution with parameters μ_x, μ_y, σ_x^2, σ_y^2, $\rho = \beta\sigma_x/\sigma_y$. It is further assumed that pairs of values (X, Y) obtained at different times are independent. Let X denote the discharge for the stream with the long record. We are concerned with estimation of μ_y and σ_y^2. The data given are pairs of observations

$$(X_1, Y_1), \ \ldots \ , (X_{n_1}, Y_{n_1})$$

for the period covered by the short record, and n_2 additional values

$$X_{n_1+1} \ , \ \ldots \ , X_{n_1+n_2}$$

from the long record.

The "synthetic values" of Y are estimated from a regression equation fitted to the n_1 paired observations.

$$\hat{Y}_{n_1+j} = \bar{Y}_1 + b(X_{n_1+j} - \bar{X}_1), \ j = 1, \ldots, n_2,$$

where

$$n_1 \bar{X}_1 = X_1 + \ldots + X_{n_1} ,$$

$$n_1 \bar{Y}_1 = Y_1 + \ldots + Y_{n_1} ,$$

$$b = \sum_{i=1}^{n_1} (X_i - \bar{X}_1)(Y_i - \bar{Y}_1) \Big/ \sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2 .$$

The mean μ_y is estimated by the mean of observed and synthetic values of Y combined,

$$U = \bar{Y}_1 + \frac{n_2}{n_1 + n_2} b(\bar{X}_2 - \bar{X}_1) ,$$

where

$$n_2 \bar{X}_2 = X_{n_1+1} + \ldots + X_{n_1+n_2} .$$

The variance of U has been obtained by Thomas,(*) who has discussed the properties of the estimator U with reference to values of n_1, n_2, and ρ. The purpose of this note is to investigate some possible estimators for σ_y^2.

(*) H. A. Thomas, Jr., "Correlation Techniques for Augmenting Stream Runoff Information", manuscript.

3. Some "Natural" Estimators for σ_y^2

First, three functions of the observations are defined.

(3.1) $$S_1^2 = \sum_{i=1}^{n_1} (Y_i - \bar{Y}_1)^2 ,$$

(3.2) $$S_2^2 = \sum_{j=1}^{n_2} (\hat{Y}_{n_1+j} - \hat{\bar{Y}}_2)^2$$

$$= b^2 \sum_{j=1}^{n_2} (X_{n_1+j} - \bar{X}_2)^2 ,$$

(3.3) $$S_3^2 = \sum_{i=1}^{n_1} (Y_i - U)^2 + \sum_{j=1}^{n_2} (\hat{Y}_{n_1+j} - U)^2 ,$$

where

$$n_2 \hat{\bar{Y}}_2 = \hat{Y}_{n_1+1} + \ldots + \hat{Y}_{n_1+n_2} .$$

Three estimators which seem to be likely candidates are as follows.

(3.4) $$T_1 = S_1^2/(n_1-1)$$

(3.5) $$T_2 = (S_1^2 + S_2^2)/(n_1 + n_2 - 2)$$

(3.6) $$T_3 = S_3^2/(n_1 + n_2 - 1)$$

The first of these, T_1, is the usual unbiased estimator based on the observed values of Y. T_3 is the estimator which would be calculated if the fact were ignored that some of the Y values were calculated. T_2 provides an alternative way of combining observed and calculated Y values.

Observe that there is a relation among these estimators, since

(3.7) $$S_3^2 = S_1^2 + S_2^2 + \frac{n_1 n_2}{n_1+n_2} b^2 (\bar{X}_2-\bar{X}_1)^2 .$$

Each of the estimators T_2, T_3 is biased, tending to give low values for the variance of Y. It will be seen, however, that they can be preferable to T_1 for sufficiently large ρ.

(3.8) $$E\,T_1 = \sigma_y^2 ,$$

(3.9) $$E\,T_2 = \sigma_y^2 - \frac{(n_2-1)(n_1-4)}{(n_1+n_2-2)(n_1-3)}\,(1-\rho^2)\sigma_y^2 ,$$

(3.10) $$E\,T_3 = \sigma_y^2 - \frac{n_2(n_1-4)}{(n_1+n_2-1)(n_1-3)}\,(1-\rho^2)\sigma_y^2 .$$

In order to make comparisons among these estimators, the variance of T_1 and the mean-squared-errors of T_2 and T_3 were calculated. These are given in formulas (3.11) - (3.13).

(3.11) $$\mathrm{Var}\,(T_1) = 2\,\sigma_y^4\,/(n_1-1) ,$$

(3.12) $$\mathrm{MSE}\,(T_2) = \mathrm{Var}\,(T_1) + \frac{(n_2-1)}{(N-2)^2}\,\sigma_y^4\Big[2A + (n_2+1)B + (N-2)C - (n_1+1)(2n_1+n_2-3)/(n_1-1)\Big] ,$$

(3.13) $$\mathrm{MSE}\,(T_3) = \mathrm{Var}(T_1) + \frac{n_2}{(N-1)^2}\,\sigma_y^4 \Big[2A + (n_2+2)B + (N-1)C - (n_1+1)(2n_1+n_2-2)/(n_1-1) \Big] ,$$

where $N = n_1 + n_2$ and

(3.14) $$A = (n_1-1)\rho^4 + (n_1+4)\rho^2(1-\rho^2) + \frac{n_1+1}{n_1-3}(1-\rho^2)^2 ,$$

(3.15) $$B = \rho^4 + \frac{6}{n_1-3}\rho^2(1-\rho^2) + \frac{3}{(n_1-3)(n_1-5)}(1-\rho^2)^2 ,$$

(3.16) $$C = 2\,\frac{n_1-4}{n_1-3}(1-\rho^2) .$$

For further abbreviation, Q_k will denote the quantity in square brackets in the expression for $\mathrm{MSE}(T_k)$, $k = 2,3$.

The "information" ratios are the reciprocals of the following.

(3.17) $$\frac{\mathrm{MSE}(T_2)}{\mathrm{Var}(T_1)} = 1 + \frac{(n_1-1)(n_2-1)}{2(N-2)^2}\,Q_2 .$$

(3.18)
$$\frac{MSE(T_3)}{Var(T_1)} = 1 + \frac{(n_1-1)n_2}{2(N-1)^2} Q_3 .$$

The notation

$$I_k(\rho, n_1, n_2) = \frac{Var(T_1)}{MSE(T_k)} , \quad k = 2,3 ,$$

will be used.

4. Properties of the Information Ratios

The following general properties hold for the two information ratios. (Where the subscript is dropped, the same statement holds for both cases.)

(4.1)
$$I_2 (1, n_1, n_2) = 1 + (n_2 -1)/(n_1 -1).$$

(4.2)
$$I_3 (1, n_1, n_2) = 1 + n_2/(n_1-1) .$$

(4.3)
$$I (0, n_1, n_2) < 1 \text{ for all } n_1, n_2 .$$

(4.4)
$$I (0, n_1, n_2) \sim 1/n_1 \text{ as } n_1 \longrightarrow \infty ,$$

with the ratio n_2/n_1 held fixed, or with the difference (n_2-n_1) fixed.

$$\rho_o \longrightarrow 1 \quad \text{as} \quad n_1 \longrightarrow \infty \;, \tag{4.5}$$

with n_2/n_1 fixed, where ρ_o is the value of ρ for which $I\ (\rho,\ n_1,\ n_2) = 1$.

5. Conclusions

From (4.4) and (4.5) it is clear that T_1 would be the preferred estimator for very large n_1, and even for moderately large n_1 if the value of ρ is not believed to be very close to unity.

For the numerical values of n_1, n_2 which would apply in the case of discharge records, the properties of T_2 and T_3 are essentially indistinguishable. T_3 would probably be preferred on grounds of convenience, if either were to be used.

The table below shows how large ρ would have to be in order that T_2 or T_3 be as good as T_1 in the sense $I\ (\rho,\ n_1,\ n_2) = 1$.

Values of ρ for which $I(\rho, n_1, n_2) = 1$

$N = n_1 + n_2$	n_1	ρ_o
30	15	.8
	20	.7
40	15	.8
	20	.8
	25	.8
	30	.8
360	180	.9

USCOMM-NBS-DC